A Blending of Computer-Based Assessment and Performance-Based Assessment: Multimedia-Based Performance Assessment (MBPA). The Introduction of a New Method of Assessment in Dutch Vocational Education and Training (VET).

Sebastiaan de Klerk[*], Theo J.H.M. Eggen[**], Bernard P. Veldkamp[***]

[*] Sebastiaan de Klerk, ECABO/University of Twente, PO Box 1230, 3800 BE Amersfoort (The Netherlands), s.dklerk@ecabo.nl.

[**] Theo J.H.M. Eggen, Cito/University of Twente, PO Box 1034, 6801 MG Arnhem (The Netherlands), theo.eggen@cito.nl.

[***] Bernard P. Veldkamp, University of Twente, PO Box 217, 7500 AE Enschede (The Netherlands), b.p.veldkamp@utwente.nl.

Abstract

Innovation in technology drives innovation in assessment (Conole & Warburton, 2005; Bartram, 2006; Drasgow, Luecht, & Bennett, 2006; Mayrath, Clarke-Midura, & Robinson, 2012). Since the introduction of computer-based assessment (CBA), a few decades ago, many formerly paper-and-pencil tests have transformed in a computer-based equivalent. CBAs are becoming more complex, including multimedia and simulative elements and even immersive virtual environments. In Vocational Education and Training (VET), test developers may seize the opportunity provided by technology to create a multimedia-based equivalent of performance-based assessment (PBA), from here on defined as multimedia-based performance assessment (MBPA). MBPA in vocational education is an assessment method that incorporates multimedia (e.g. video, illustrations, graphs, virtual reality) for the purpose of simulating the work environment of the student and for creating tasks and assignments in the assessment. Furthermore, MBPA is characterized by a higher amount of interactivity between the student and the assessment than traditional computer-based tests. The focal constructs measured by MBPA are the same as are currently assessed by performance-based assessments. Compared to automated delivery of item-based tests, MBPA realizes the full power of ICT. In the present article we will therefore discuss the current status of MBPA, including examples of our own research on MBPA. We provide an argument for the use of MBPA in vocational education too.

Keywords: assessment in vocational education and training, performance-based assessment, computer-based assessment, multimedia-based performance assessment

A Blending of Computer-Based Assessment and Performance-Based Assessment: Multimedia-Based Performance Assessment (MBPA). The Introduction of a New Method of Assessment in Dutch Vocational Education and Training (VET).

Technological advancements continue to drive innovation in CBA. Presently, test developers already incorporate multimedia elements into their CBAs (de Klerk, 2012). Educational institutions are designing technology-based assessments and items in a growing rate (Bennett, 2002). These assessments simulate a highly contextualized environment and students are confronted with tasks that they could encounter in real life (Issenberg, Gordon, Gordon, Safford, & Hart, 2001; Ziv, Small, & Wolpe, 2000). The general rationale behind innovative assessments is that they provide more meaningful observations about student skills than traditional multiple-choice tests or performance-based assessment (PBA). For example, research has shown that immersive environments are capable of capturing observations that are not possible to capture in a conventional classroom setting (Clarke, 2009; Ketelhut, Dede, Clarke, Nelson & Bowman, 2008). In scientific literature, multimedia and even immersive virtual reality possibilities are being discussed in relation to assessment (Susi, Johanneson, & Backlund, 2007; Sliney, & Murphy, 2011; Clarke-Midura & Dede, 2010). Thus, technological progress provides opportunities for the design and development of innovative and interactive technology-based assessments. However, technological advancement in assessment is ahead of research and psychometrics.

Therefore, in this article, we present a research project that tries to fill the void between the opportunities that technology provides for assessment and the foundation of these opportunities in theory and research. We provide an argument for technology-based assessments, we discuss the current status of technology in assessment, and we present our current research on an innovative method of assessment in Dutch vocational education, called multimedia-based performance assessment (MBPA). MBPA in vocational education is an

assessment method that incorporates multimedia (e.g. video, illustrations, graphs, virtual reality) for the purpose of simulating the work environment of the student and for creating tasks and assignments in the assessment. Furthermore, it is characterized by a higher amount of interactivity between the student and the assessment than in traditional computer-based tests. Finally, the focal constructs under measurement in MBPA are the same as are currently assessed using performance-based assessment. The introduction of MBPA and technology in assessment in general is not just about doing the same things differently. It is primarily about introducing a measurement instrument to the vocational educational field that provides new and improved possibilities for measuring students' skills.

The vocational education and training (VET) sector may be strongly diversified in Europe, or even worldwide, for that matter. However, the core elements of VET have universal applications over countries. Through the use of MBPA we try to translate these elements in a computer-based assessment. First, one of the core elements in VET is that it prepares students for vocations that for the largest part emphasize manual/practical skills or procedural knowledge. Therefore, the tasks in the MBPA should be designed around these constructs. Secondly, although countries differ in extent, for some part VET is always carried out in the original vocational setting during an apprenticeship or internship. The MBPA should reflect the vocational context as if students were working in a real setting. For instance through video material or even virtual reality elements. Finally, VET is always concerned with getting students to a proficiency level with which they can act as entry employees in their vocation on the labor market. Thus, the MBPA should distinguish between students that can act as entry employees (mastery) and students that have not reached the right level of proficiency yet (non-mastery). The main goal for assessment methods in VET, therefore, is to reach high levels of predictive validity.

Unfortunately, the current assessment methods in vocational education are not sufficient for validly measuring students' skills and competencies. For example, one of the most common assessment methods in vocational education, PBA, is prone to several sources of measurement error. In PBA, the generalizability of scores may be impaired by several factors; due to task and rater error (Dunbar, Koretz, & Hoover, 1991), due to administration occasion error (Cronbach, Linn, Brennan, & Haertel, 1997) or due to assessment method error (Shavelson, Baxter, Gao, 1993). Furthermore, it is difficult to standardize the assessment setting and PBAs are time consuming, expensive, and logistically challenging (Lane & Stone, 2006).

Technology offers interesting opportunities to create assessments that are capable of improved measurement of the same constructs that are now measured with a PBA. Research should point out whether innovative assessments can fulfil this promise or not. Therefore, in the current article, we present an argument that, based on the current status of technology in assessment, justifies a solid investigation into the use of MBPA in vocational education. The presented argument is a first effort to bring technological advancements in assessment and research together. To further specify our argument we also present a pilot example of our first attempt to create an MBPA for a vocation that is now assessed with a PBA. Next, we discuss the background context of our research project, and will then turn the discussion to our argument for the use of MBPA in vocational education.

Assessment in Vocational Education and Training: Challenges and Concerns

Student learning in vocational education generally revolves around acquiring complex and integrated knowledge, skill, and attitude constructs which are often referred to as 'competency' (Klieme, Hartig, Rauch, 2008). Of course, competency is not directly observable in students (Grégoire, 1997), thus to make statements about student competencies we have to rely on indirect reasoning from evidence that we collect in an assessment setting.

In the assessment setting the student is confronted with assignments or tasks that require responses or behaviors. Subject matter experts (SMEs) and assessment experts together develop a model that reflects the degree to which the student has mastered a competency based on the performance of the student in the assessment setting. Student learning in vocational education is becoming more and more focused on mastery of competency. Id est, vocational education has shifted from a traditional testing culture to an assessment culture (Baartman, 2008). In practice, this resulted in an increasing emphasis on a comprehensive alternative form of assessment, generally referred to as ‘authentic assessment’, ‘alternative assessment’, or ‘performance(-based) assessment’ (Linn, Baker, & Dunbar, 1991; Marzano, Pickering, & McTighe, 1993; Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson, Baxter, & Pine, 1992; Sweet & Zimmermann, 1992). Although multiple-choice tests and other assessment methods are still used, performance-based assessment (PBA) is now the most pervasive assessment method in vocational education (Segers, 2004; Baartman, 2008; Dierick & Dochy, 2001; Van Dijk, 2010).

Traditional paper-and-pencil tests are not capable of capturing the complex and integrated constructs assessed in vocational education but are still used when the acquisition of knowledge is tested. Additionally, PBA has high face validity when competency is the focal construct of assessment. For example, compared to paper-and-pencil tests most people would choose PBA to be the prevailing instrument to measure vocational competencies. Although PBA seems to be a promising method of assessment in vocational education it imposes some serious challenges and concerns upon its developers and users. Before the challenges and concerns of PBA can be discussed it should be noted that there exist multiple types of PBA. Roelofs and Straetmans (2006) discuss three types of PBAs: hands-on, simulation-based, and hands-off.

Hands-on PBAs are assessments that take place during students' work placement. The evidence of competency is collected during the observation of students performing a vocation in real life. Simulation-based PBAs are assessments that take place in more or less standardized and reconstructed settings (e.g. in school). Simulations simplify, manipulate or remove parts of the natural job environment and students are cognizant of the fact that the situation is not a real-world setting. Hands-off PBAs are paper-based assessments in which students are confronted with hypothetical vocational situations. Subsequently, the students are asked how they would or should react in these situations. In the current article we will only refer to the first two types because those are used in vocational education.

Both types of PBA have their pros and cons. For example, hands-on PBAs are generally considered to be very authentic, which would increase their validity. However, they are also very difficult to standardize and sometimes students are not allowed to carry out specific tasks because of risk. Simulation-based assessments, on the contrary, can be more standardized measures of competency but are less authentic. In practice, hands-on as well as simulation-based PBAs are prone to several measurement issues and practical concerns. In the table below, some (but not all) characteristics of hands-on and simulation-based PBAs are presented. We consider these characteristics to be the most important characteristics of PBA and the most important for the current discussion.

Insert Table 1 about here

PBAs are generally characterized by flexible open ended tasks. Students are required to construct original responses, which results in a unique assessment for every student, and generally more than one correct outcome. Above that, the assessment setting is called authentic and meaningful because it resembles a real-life setting and students can better

associate with the tasks compared to traditional measurement (Gulikers, Bastiaens, & Kirschner, 2004).

This leads to the first, and one of the most important measurement concerns of PBA: the interaction between standardized measurement on the one hand, and an authentic assessment setting on the other hand. As can been seen in Table 1, the typical hands-on PBA takes place in an authentic assessment setting, but lacks standardized tasks. Possible concerns in standardization could be detected using generalizability theory. Generalizability theory allows for the estimation of multiple sources of error in measurement (Brennan, 1983, 200, 2001; Cronbach, Gleser, Nanda & Rajaratnam, 1972; Shavelson & Webb, 1991).

In generalizability theory, variance components can be estimated for each facet of the assessment and the interactions between facets (Lane & Stone, 2006). Facets are, for example, the tasks, raters, and occasion. These variance components indicate to what extent the facets in the assessment cause measurement error. Furthermore, generalizability theory provides coefficients that are used to examine how well the assessment scores generalize to the larger construct domain. Poor standardization in the tasks, raters or occasion may translate into low generalizability coefficients and construct irrelevant variance. This is exactly what has been found in empirical research on measurement error in PBA. Measurement error originates due to the selected tasks in the assessment (Baxter, Shavelson, Herman, Brown, & Valdez, 1993; Gao, Shavelson, & Baxtor, 1994), due to the use of raters in the assessment (Breland, 1983; Dunbar, Koretz, & Hoover, 1991; van der Vleuten & Swanson, 1990), due to the occasion (Cronbach, Linn, Brennan, & Haertel, 1997), or a combination of those (Shavelson, Ruiz-Primo, & Wiley, 1999).

The second concern related to PBA in vocational education is the use of raters. Several studies have shown that rater induced error and inter-rater reliability varies considerably between different performance-based assessments (Clauser, Clyman, & Swanson, 1999). For

example, Dunbar, Koretz, and Hoover (1991) found reliability levels ranging from .33 to .91, and van der Vleuten and Swanson (1990) reported reliability coefficients in the range of .50 to .93. The dispersion of these numbers shows that rater induced error is a significant factor concern. Rater induced error has already been exemplified in an excellent manner by Edgeworth in 1888 (cited by Bejar, Williamson, & Mislevy, 2006): *... let a number of equally competent critics independently assign a mark to the (work). ... even supposing that the examiners have agreed beforehand as to ... the scale of excellence to be adopted .. there will occur a certain divergence between the verdicts of competent examiners. (p.2).*

Moreover, many rater effects that influence scores have been described in literature (Eckes, 2005; Dekker & Sanders, 2008). These effects unintentionally affect the scores of students and cause construct irrelevant variance. Hence, the variance that is created in the ratings of students through the rater effects threatens the validity of the assessment (Messick, 1989, 1995; Weir, 2005). The most well-known example of a rater effect is the halo effect (Thorndike, 1920). The halo effect is a cognitive bias that influences our judgment of particular behavior of students based on the overall impression of the student to be judged. The training of raters and raising their conscience on these effects is found to reduce the effects' influence (c.f. Wolfe & McVay, 2010). However, rater effects are inevitable, simply because the raters are human.

The third concern is the representativeness of PBA. Fitzpatrick and Morrison discuss comprehensiveness and fidelity as two aspects of the representativeness of PBA (1971, p.240): *comprehensiveness, or the range of different aspects of the situation that are simulated, and fidelity, the degree to which each aspect approximates a fair representation of that aspect in the criterion.* That is, within each assessment domain there is a whole range of possible tasks to include in the assessment. However, which tasks are included in the final assessment may influence students' performance. Ideally, no matter which tasks selected, they

should all result in the same statements about students' mastery of skill (Straetmans & Van Diggele, 2001). In reality, students usually perform a limited amount of tasks in a PBA, because it is costly and logistically challenging to have students perform many tasks in many situations. This results in insufficient representativeness of the PBA. Poor representativeness combined with limited standardization and rater induced error are the main sources of low reliability of PBAs.

The fourth and final concern, feasibility, is not a measurement concern, but may result in diminished measurement quality. In general, PBA is considered an inefficient type of assessment. Efficiency in assessment can be described along several dimensions. The most important are time, costs, and logistics. PBAs are often time consuming, costly, and logistically challenging to design and develop. To retain some efficiency vocational schools may cut back on resources used for the assessment (i.e. money or time) or on technical aspects of the assessments (i.e. psychometric evaluation), thereby reducing the validity and overall quality of the assessment (Shavelson, Baxtor, & Gao, 1993; Cronbach, Linn, Brennan, & Haertel, 1997; Shavelson, Ruiz-Primo, & Wiley, 1999; Webb, Schlackman, & Sugrue, 2000; Haertel, Lash, Javitz, & Quellmalz, 2006).

In summary, the concerns described above indicate that it is very difficult to design and develop PBAs that provide valid results about student competency. This has led Kane (1992) to a firm conclusion about the development of PBA: *basically you can't win.* At the least, it requires very careful and thoughtful development, extensive rater training, and detailed psychometric analysis to be comfortable about its functioning. These things are, unfortunately, often not possible when budget and time are considered. Technological advancement in assessment may provide the potential to realize assessments that do provide valid and reliable statements about student mastery of competency combined with practical and cost efficient development and administration.

A Rationale for Multimedia-Based Performance Assessment

Multimedia-Based Performance Assessment is grounded in the realization of the full power ICT provides. The general rationale behind MBPA is: (1) that it enables improved measurement of competency compared to PBA because MBPA might be less prone to several of the measurement concerns discussed above, and (2) that it is a more efficient assessment method than PBA.

MBPA has higher standardization than PBA and still retains authenticity (Clarke, 2009; Schoech, 2001; Bakx, Sijtsma, Van Der Sanden, & Taconis, 2002). Furthermore, MBPAs provide the possibility to improve the representativeness of the assessment. For example, students may be presented with a multitude of situations and tasks in the MBPA compared to a PBA. To have students perform multiple tasks in multiple situations enhances the reliability of the instrument. MBPA provides the possibility to confront students with highly contextualized critical situations. Thereby, the two aspects of the representativeness of the measure, namely comprehensiveness and fidelity, are improved considerably using MBPA.

Another advantage of MBPA is that raters are ruled out of scoring. As mentioned above, raters are one of the most important sources of measurement error in PBA. MBPA enables test developers and administrators to automatically score student performance. Rater effects, and subsequently rater induced measurement error, can thereby be diminished.

Finally, compared to PBA, MBPA has higher efficiency. The assessment is administered virtually rather than physically. That is, physical situations where students, raters, and possible actors have to meet to perform the PBA are not needed in MBPA. Furthermore, vocational schools can simultaneously administer the MBPA in large groups. Compared to PBA, which is individually administered, MBPA is more efficient on this aspect.

To summarize, the rationale behind MBPA is not just to replicate what PBA is also capable of, but to add new features to the measurement spectrum in vocational education. Or, to quote Thornburg (1999): *The key idea to keep in mind is that the true power of educational technology comes not from replicating things that can be done in other ways, but when it is used to do things that couldn't be done without it. (p. 7).* We have hypothesized that MBPA might perform better on standardization, representativeness, reliability, rater effects, and efficiency than PBA.

Research on Innovations in Assessment

Research on innovations in computer-based assessment covers a wide variety of topics, from using technology in testing to detect potential fraud (Wollack & Fremer, 2013) on the one end to highly interactive virtual reality assessments on the other end (Clarke-Midura & Dede, 2010). The most recent innovation that is now widely implemented and is most relevant for the current discussion is the introduction of so-called innovative (or alternative) item types. Scalise and Gifford (2006) present an overview of different innovative item types that have emerged and become available for test developers. Innovative item types sometimes incorporate multimedia (e.g. graphs or illustrations) and research has shown that these items provide test developers with the opportunity to test specific aspects of student learning (e.g. application of knowledge, inquiry) that are impossible to test with traditional multiple-choice tests (Scalise & Gifford, 2006).

Current technological progress creates unparalleled possibilities for assessment and announces a whole new era of assessments to look forward to. For example, researchers are now starting to introduce immersive virtual environments into the educational measurement field (Clarke, 2009). Originally stemming from e-learning applications (see for example Monahan, McArdle, & Bertolotto, 2008), these virtual environments simulate a more or less real-world environment which often requires highly interactive operating. Students can

perform a wide variety of tasks and objectives within the virtual environment and the computer automatically records, logs, and scores student behavior.

Research on the use of multimedia in assessment is done in the field of organizational psychology as well. For example, Oostrom, Born, Serlie, and Van der Molen (2011) have done research on the use of a multimedia situational test in which the test items are presented as video clips. Furthermore, the same group of researchers (2010) has also experimented with an innovative open-ended multimedia test in which the test takers responses are recorded with a webcam. Finally, serious gaming and assessment is another multimedia influenced topic of research. For example, researchers and practitioners have designed and tested virtual manager games to assess test takers competencies (Chang, Lee, Ng, & Moon, 2003).

It is important to note that these innovations do not solely result from innovative technology. Technology enables researchers and test developers to design innovative CBAs, but technology does not determine the success of assessment innovations (Williamson, Bejar, & Mislevy, 2006). It is the combination of technology and structured design, grounded in everything we know about assessment that results in solid and coherent assessments. Williamson et al. subtly remark that technological advances have even outpaced general assessment methods for the interpretation of scores that result from innovative CBAs. Of course, the challenge lies not in developing richly designed, highly contextualized and magnificently looking multimedia-based assessments, however the challenge lies in having them function psychometrically as well.

Introducing Multimedia-Based Performance Assessment in Vocational Education and Training

To illustrate the added value of MBPA in VET we now present a pilot example of an MBPA that we are designing and developing. The MBPA is used to assess students' skills and abilities for a specific vocational education: *safety guard for confined spaces*. Currently,

students that pass a PBA become certified safety guards. A safety guard for confined spaces ensures that confined space work is carried out responsibly and safely by workers. For example, by doing job safety analyses and maintaining an adequate communications system. Our goal is to develop an MBPA that is capable of capturing students' mastery of skills in a way that is more valid and reliable than the current PBA. The first research question therefore is:

RQ1. Based on psychometric and empirical comparison; to what extent does the MBPA perform better than the current PBA?

The design and development will take place according to a developmental framework for MBPA in VET that is currently under review for publication (De Klerk, Veldkamp, & Eggen, *submitted for review*). During the research project we will mainly focus on studying the measurement properties (validity and reliability) and the efficiency of MBPA. The second research question that we try to answer in this research project therefore is:

RQ2. Using the framework referred to above; can we construct an MBPA that provides valid and reliable inferences about students' mastery of skills to be certified as a safety guard confined spaces?

Finally, based on the results of the current research project and gained experience the third research question is:

RQ3. Is MPBA both an efficient and effective assessment method in VET?

Multimedia-Based Performance Assessment: An Example

The MBPA we are designing and developing is a simulation of the real job of a safety guard confined spaces and is based upon real work processes. Because of the simulative element we can quickly change the tasks or situations that the safety guard is virtually confronted with. The first pilot version of the MBPA has already been created. Using video

clips in which a real safety guard and context is presented we provide students with a virtual environment in which they can already perform several tasks.

Insert Figure 1 about here

In the assessment, students follow the safety guard and a worker performing their tasks (guarding and cleaning a confined space) and students are required to intervene when they observe incorrect behavior of the worker performing in and around the confined space and of the safety guard self. For example, safety guards have to determine the optimal escape route in case of a hazard or a factory alarm. One aspect of an optimal escape route is the wind, and when students observe the safety guard determining the escape route incorrectly they can intervene by pressing the stop button, see Figure 2. Also, students have the opportunity to study the work permit during the assessment as can be seen in Figure 3. For example, to see which gasses or substances have been in the confined space or to identify possible hazards.

Insert Figure 2 about here

Insert Figure 3 about here

Students are introduced to the assessment via a video that explains what they are going to see and what their options and tools are in the assessment. In the pilot version we log and score all interventions made during the assessment. When students intervene, a new window pops up in which they can type the incorrect behavior they have observed (see Figure 4). Key terms are possible to score (e.g. “ear protection” in the example) as well as sentences, but the results provided by the pilot version still need to be scanned by a rater. Thus, students have to

observe and decide about erratic behavior displayed by either the worker or the safety guard and intervene when they do see that happening. All interventions and student reactions are recorded and provide an observation about students' mastery of skills to work as a safety guard.

Insert Figure 4 about here

Method

With co-funding from the Foundation Cooperation for Safety (SSVV) we are currently working on the design and development of an expanded version of the MBPA pilot version for safety guards confined spaces. The new assessment incorporates more multimedia elements and a higher amount of interactivity between student and assessment. Above that, the MBPA should be fully independent of raters; all actions of students have to be logged and scored automatically. Of course, the point of departure for implementing technological improvements in an MBPA should always be improved measurement of students' mastery of skills. That is, with these new technologies we try to produce better observations about student learning than is possible with traditional measurement methods (e.g. PBA). For example, the PBA is limited in the amount of situations and tasks a student can perform but in the MBPA we can confront students with a multitude of tasks and situations In addition, we can provide students with tools (e.g. a work permit, communication set, and measurement instruments) and continually update their status and the information in the MBPA as they progress through the MBPA. Furthermore, we are able to log and save everything that the student does in the virtual environment for later (psychometric) analysis.

To answer the research questions posed above, we plan to conduct several studies based upon the MBPA for safety guards confined spaces that we are developing. First, we will

be studying the psychometric functioning of the assessment using the versatile data that the MBPA produces. One of the challenges in using MBPA in certification settings lays not so much in the design and development of MBPA, but in the psychometric analysis of the data that it produces. For example, which responses or variables in the assessment provide evidence about students' mastery of skills? We will fit different psychometric models to the data (e.g. IRT models and classical test theory), and we will determine the MBPA's reliability.

Secondly, we want to perform a comprehensive validity study based on Kane's (1992) argument-based approach to validation. Validity is probably the most central concept in assessment (Messick, 1989). The argument-based approach to validation emphasizes the evaluation of the plausibility of the various assumptions and inferences involved in interpreting assessment observations as a reflection of students' mastery of skills. Of course, psychometric functioning and reliability are part of the assessments' overall validity but we also want to include an empirical comparison of the MBPA and PBA as a validity argument. In an empirical study, students will either first do the PBA and then the MBPA or the other way around. We will then analyze results from both assessments and report our findings in a future article.

Conclusion

Technology provides unparalleled opportunities in assessment. Now, assessment developers, practitioners and researchers may seize the opportunity to develop and study a new type of assessment in vocational education: multimedia-based performance assessment. This is a valuable endeavor because technological possibilities are ahead of psychometrics. Furthermore, current assessment methods in VET may not be sufficient to validly and reliably measure every aspect of students' mastery of skills. We have shown that time and resources

make it difficult to design and develop PBAs that provide valid and reliable inferences. MBPA might provide a solution to this problem.

In contrast to PBA, MBPA is fully standardized which reduces measurement error in measuring students' skills. Using MBPA it is also possible to present more situations and tasks than in PBA, which implicates improved representativeness of the MBPA compared to the PBA. Improved representativeness also results from the possibility to incorporate high risk tasks and infrequent tasks in an MBPA. Furthermore, one of the major causes of construct irrelevant variance in PBA, raters, can be ruled out of the scoring process. Together, this may result in more valid and reliable MBPA scores compared to PBA scores. Finally, the feasibility of MBPA may be higher than PBA because there is no need for printed materials and personnel (actors, raters, etc.). MBPA provides the opportunity of large scale and remote administration too.

The main goal of the research project presented in this article is to investigate the overall validity, reliability and feasibility of MBPA. We are trying to answer the question whether MBPA is both an effective and efficient method of assessment in VET. Currently, we look at MBPA as a promising assessment method for the assessment programs of most qualifications in VET. However, we also expect that in a first stage MBPA will complement rather than replace the traditional measurement methods in VET. Research should first point out how to use, psychometrically speaking, the data that MBPA produces. Furthermore, validation studies are needed to determine the practical value of MBPA in an educational setting. Thus, although MBPA is full of promise, there is still a lot of work to be done before actual large-scale implementation can take place.

References

Baartman, L.K.J. (2008). *Assessing the assessment: Development and use of quality criteria for competence assessment programmes* (Doctoral dissertation, Utrecht University, The Netherlands). Retrieved from http://hdl.handle.net/1820/1555.

Bakx, A.W.E.A., Sijstma, K., Van Der Sanden, J.M.M., & Taconis, R. (2002). Development and evaluation of a student-centerd multimedia self-assessment instrument for social-communicative competence. *Instructional Science, 30,* 335-359.

Bartram, D. (2006). Testing on the internet: issues, challenges, opportunities in the field of occupational assessment. In D. Bartram and R.K. Hambleton (Eds.), *Computer-Based Testing and the Internet* (pp. 13-37). Chichester: Wiley.

Baxter, G.P., Shavelson, R.J., Herman, S.J., Brown, K.A., & Valadez, J.R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Research in Mathematics Education, 24*, 190-216.

Bejar, I.I., Williamson, D.M., & Mislevy, R.J. (2006). Human scoring. In D.M. Williamson, R.J. Mislevey & I.I Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49-81). Mahwah, NJ: Lawrence Erlbaum.

Bennett, R.E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. In D. Bartram and R.K. Hambleton (Eds.), *Compouter-Based Testing and the Internet* (pp. 201-217). Chichester: Wiley.

Breland, H. (1983). *The direct assessment of writing skill: A measurement review* (College Board Report No. 83-6). New York: College Entrance Examination Board.

Brennan, R.L. (1983). *Elements of generalizability*. Iowa City, IA: American College Testing Program.

Brennan, R.L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*, 339-353.

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Chang, J., Lee, M., Ng, K., & Moon, K. (2003). Business Simulation Games: The Hong Kong Experience. *Simulation & Gaming, 34*, 367-376.

Clarke, J. (2009). *Studying the potential of virtual performance assessment for measuring student achievement in science.* Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, CA. Retrieved September 5, 2012, from http://virtualassessment.org/publications/aera_2009_clarke.pdf

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education, 42*(3), 309-328.

Clauser, B.E., Clyman, S.G., & Swanson, D.B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement, 36*(1), 29-45.

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. *Research in Learning Technology, 13(1),* 17-31.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: John Wiley.

Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373-399.

De Klerk, S. (2012). An overview of innovative computer-based testing. In T.J.H.M. Eggen & B.P. Veldkamp (Eds.), *Psychometrics in practice at rcec* (pp. 137-150). Enschede: RCEC.

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (submitted for review). A framework for designing and developing multimedia-based performance assessment in vocational education.

Dekker, J., & Sanders, P.F. (2008). *Kwaliteit van beoordeling in de praktijk* [Quality of rating during work placement]. Ede: Kenniscentrum handel.

Dierick, S., & Dochy, F.J.R.C. (2001). New lines in edumetrics: new forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*, 307-329.

Drasgow, F., Luecht, R.M., & Bennett, R.E. (2006). Technology and testing. In R.L. Brennan (Ed.), *Educational Measurement* (pp. 471-530). Westport, CT: Praeger.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-304.

Eckes, T. (2005). Examining rater effects in TestDaf writing and speaking performance assessments: A many-facet Rasch analysis. *Language assessment quarterly, 2*(3), 197-221.

Edgeworth, F.Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*, 599-635.

Fitzpatrick, R., & Morrison, E.J. (1971). Performance and product evaluation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 237-270). Washington DC: American Council on Education.

Gao, X., Shavelson, R.J., & Baxter, G.P. (1994). Generalizability of large-scale performance assessments in science. Promises and problems. *Applied Measurement in Education, 7*, 323-334.

Grégoire, J. (1997). Diagnostic assessment of learning disabilities. From assessment of performance to assessment of competence. *European Journal of Psychological Assessment, 13*(1), 10-20.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52(3),* 67-86.

Haertel, E.H., Lash, A., Javitz, H., & Quellmalz, E. (2006). An instructional sensitivity study of science inquiry items from three large-scale science examinations. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Issenberg, S.B., Gordon, M.S., Gordon, D.L., Safford, R.E, & Hart, I.R. (2001). Simulation and new learning technologies. *Medical Teacher, 23*(1), 16-23.

Kane, M.T. (1992). The assessment of professional competence. *Evaluation & the health professions, 15*(2), 163.

Ketelhut, D. J., Dede, C., Clarke, J., Nelson, B., & Bowman, C. (2007). Studying situated learning in a multi-user virtual environment. In E. Baker, J. Dickieson, W. Wulfeck, & H. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 37–58). Mahwah, NJ: Lawrence Erlbaum.

Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In E. Klieme, J. Hartig, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 3-22). Göttingen: Hogrefe.

Lane, S, & Stone, C.A. (2006). Performance assessment. In R.L. Brennan (Ed.), *Educational Measurement* (pp. 387-431). Westport, CT: Praeger.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex performance assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model*. Alexandria, VA: Association for Supervision and Curriculum Development.

Mayrath, M.C., Clarke-Midura, J., & Robinson, D.H. (2012). Introduction to technology-based assessments for 21st century skills. In M.C. Mayrath, J. Clarke-Midura, D.H. Robinson & G. Schraw (Eds.), *Technology-based assessments for 21st century skills* (pp. 1-11). Charlotte, NC: Information Age.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Monahan, T., McArdle, G., & Bertolotto, M. (2008). Virtual reality for collaborative e-learning. *Computers & Education, 50*, 1339-1353.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology, 19*(5), 532-550.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology, 10*(2), 78.

Roelofs, E.C., & Straetmans, G.J.J.M. (Eds.) (2006). *Assessment in actie* [Assessment in action]. Arnhem: Cito.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1), 41-53.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6). Retrieved [March 20, 2012] from http://www.jtla.org.

Schoech, D. (2001). Using video clips as test questions: The development and use of a multimedia exam. *Journal of Technology in Human Services, 18*(3-4), 117-131.

Segers, M. (2004). Assessment en leren als twee-eenheid: onderzoek naar de impact van assessment op leren [the dyad of assessment and learning: a study of the impact of assessment on learning]. *Tijdschrift voor Hoger Onderwijs, 22*, 188-220.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory*. Newbury Park, CA: Sage.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

Shavelson, R.J., Ruiz-Primo, M.A., & Wiley, E. (1999). Note on sources of sample variability in science performance assessments. *Journal of Educational Measurement, 36*(1), 56-69.

Sliney, A., & Murphy, D. (2011). Using Serious Games for Assessment. In M. Ma, A. Oikonomou & L. C. Jain (Eds.), *Serious Games and Edutainment Applications* (pp. 225-243). London: Springer.

Straetmans, G.J.J.M., & van Diggele, J.B.H. (2001). *Anders opleiden, anders toetsen* [Different instruction, different assessment]. BVE-brochurereeks: Perspectief op Assessment, deel 1 [BVE-brochure series: Perspective on Assessment, part 1]. Arnhem: Cito.

Susi, T., Johannesson, M., & Backlund, P. (2007). Serious games – an overview [Technological Report]. Retrieved from http://www.his.se/PageFiles/10481/HS-IKI-TR-07-001.pdf.

Sweet, D., & Zimmermann, J. (1992). Performance Assessment. *Education Research Consumer Guide*(2), 2-5.

Thornburg, D.D. (1999). *Technology in K-12 education: Envisioning a new future*. Retrieved October 16, 2012, from http://www.edtech.ku.edu/resources/portfolio/examples/nets/Miller/www.air.org/forum/Thornburg.pdf.

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29.

Van Dijk, P. (2010). *Examinering in de beroepspraktijk* [Assessment in vocational practice]. Amersfoort: ECABO.

Van der Vleuten, C.P.M., & Swanson, D.B. (1990). Assessment of clinical skills with standardized patients: The state of the art. *Teaching and Learning in Medicine, 2,* 58-76.

Webb, N.M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education, 13*(3), 277-301.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach.* Houndmills, England: Palgrave Macmillan.

Williamson, D.M., Bejar, I.I., & Mislevy, R.J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D.M. Williamson, R.J. Mislevey & I.I Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 1-13). Mahwah, NJ: Lawrence Erlbaum.

Wolfe, E.W., & McVay, A. (2010). *Rater effects as a function of rater training context.* Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/6435A0AF-0C12-46F7-812E-908CBB7ADDFF/0/RaterEffects_101510.pdf

Wollack, J.A., & Fremer, J.J. (Eds.) (2013). *Handbook of test security*. New York, NY: Routledge.

Ziv, A., Small, S.D., & Wolpe, P.R. (2000). Patient safety and simulation-based medical education. *Medical Teacher, 22*(5), 489-495.

Table 1

Types of Performance-Based Assessment and Corresponding Characteristics

	Type	
Characteristic	Hands-on	Simulation
Standardization	-	±
Authenticity	+	±
Rater induced error	+	+
Representativeness	-	-
Feasibility	-	±
Reliability	-	±

Note. + = PBA type scores high on particular feature, ± = PBA type scores neither high nor low on particular feature, - = PBA type scores low on particular feature. Table 1 is a rough delineation of corresponding characteristics for the types of PBA. The table is based on a synthesis of this paper's literature.

Figure 1. Safety guard for confined spaces in an authentic work environment.

Figure 2. Safety guard determines optimal escape route.

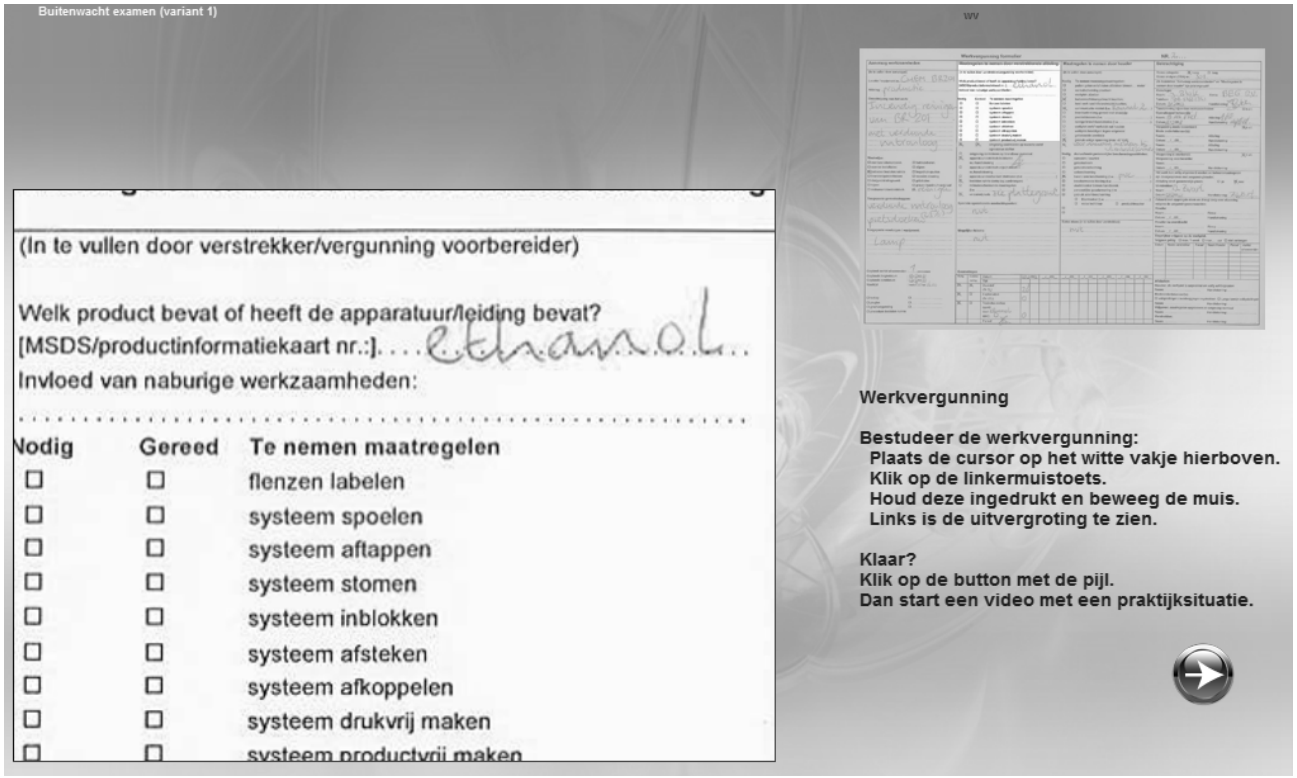

Figure 3. Students can open the work permit during the assessment.

Figure 4. Intervention made by student.